Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2005 Proceedings

Americas Conference on Information Systems (AMCIS)

2005

A Fuzzy Mining Algorithm for Association-Rule Knowledge Discovery

Lingling Zhang
Chinese Academy of Sciences, zhangll@gscas.ac.cn

Yong Shi
University of Nebraska at Omaha, yshi@gscas.ac.cn

Xinhua Yang
Alliance - PKU Management Consulting Company Limited, yangxinhua@allpku.com

A Fuzzy Mining Algorithm for Association-Rule Knowledge Discovery

Lingling Zhang
Chinese Academy of Sciences Research Center on Data Technology and Knowledge Economy, Beijing (100039), China
Management School of Graduate University of Chinese Academy of Sciences, Beijing (100080), China
zhangll@gscas.ac.cn

Yong Shi
Chinese Academy of Sciences Research Center on Data Technology and Knowledge Economy, Beijing (100039), China
College of Information Science and Technology, University of Nebraska at Omaha, NE 68182. U.S.A.
yshi@gscas.ac.cn

Xinhua Yang
Alliance- PKU Management Consulting Company Limited, Beijing (100080), China
yangxinhua@allpku.com

ABSTRACT

ABSTRASCT

Due to increasing use of very large database and data warehouses, discovering useful knowledge from transactions is becoming an important research area. On the other hand, using fuzzy classification in data mining has been developed in recent years. Hong and Lee proposed a general learning method that automatically derives fuzzy if-then rules and membership functions from a set of given training examples using a decision table. But it is complex if there are many attributes or if the predefined unit is small. Hong and Chen improve it by first selecting relevant attributes and building appropriate initial membership functions. Based on Hong's heuristic algorithm of membership functions and *Apriori* approach, we propose a fuzzy mining algorithm to explore association rules from given quantitative transactions. Experimental results on Iris data show that the proposed algorithm effectively induces more association rules.

Keywords

Data mining, Association-Rule, fuzzy classification.

INTRODUCTION

Knowledge Discovery in Database (KDD) is an important process of Knowledge Management (KM). As Information Technology (IT) progresses rapidly, many enterprises have stored large amounts of transaction data, how to extract available implicit knowledge to aid decision making from it has became a new and challenging task. Vigorous efforts have thus been devoted to designing efficient mechanisms for mining information and knowledge from large database. As a result, data mining first proposed by Agrawal et.al. in 1993[1]. Deriving association rules from transaction database is most commonly seen in data mining [1- 4,17].

Fuzzy systems that can automatically derive fuzzy if–then rules from numeric data have been developed.[5-7]Hong and Lee proposed a general learning method for automatically deriving fuzzy if-then rules and membership functions from a set of given training examples by merging the decision tables and membership functions[8]. But, it is complex if there are many attributes or if the predefined unit is small. Hong and Chen improved it by first selecting relevant attributes and building appropriate initial membership functions. These attributes and membership functions also are used in a decision table to derive final fuzzy if-then rules and membership functions [5]. Hong et al also proposed a fuzzy mining algorithm for managing quantitative data [5].
Based on Hong's heuristic algorithm of membership functions and *Apriori* approach, we propose a fuzzy mining algorithm to find association rules from given quantitative transactions. Experimental results on Iris data show that the proposed algorithm effectively induces more association rules.

In the following sections, the paper is organized as follows: In section 2, Fuzzy method in mining association rules is reviewed. In section 3, the architecture of the proposed learning algorithm and details of the proposed learning algorithm are illustrated. In section 5 experiments to verify the accuracy of proposed learning algorithm are stated. Finally, the conclusion is given in section 5.

REVIEW OF DISCOVERING FUZZY ASSOCIATION-RULES USING FUZZY METHODS

Data mining is the process of extracting previously unknown and potentially useful hidden predictive information from large amounts of data. Discovering association rules is one of the several data mining techniques described in the literate. Association-rule mining is a two –step process [9].
(1) Find all frequent item-sets: By definition, each of these items will occure at least as frequently as a pre-determined minimum support count ($MinSup$).

(2) Generate strong association rules from the frequent item-sets: By definition, these rules must satisfy minimum support ($MinSup$) and minimum confidence ($MinConf$)

We can see that the second step is easy while the first is critical and is the core of the association-rule mining algorithm.

Data mining first was proposed by Agrawal et.al. in 1993. Agrawal et. al [1] proposed a method to find the frequent item-sets. Subsequently, Agrawal et.al [10] also proposed the Apriori algorithm. However, these algorithms are proposed according to Boolean attribute, typical Apriori, and got Boolean association-rule after mining. But in practice, besides Boolean attribute, which only includes two values, most are multiple valued attributes. As Figure 1 shows, multiple valued attributes can be divided into Classification Attribute and Quantitative Attribute. The range set of classification attribute is discrete, which can be translated into Boolean association-rule mining, however the quantitative attribute's range set is continuous.

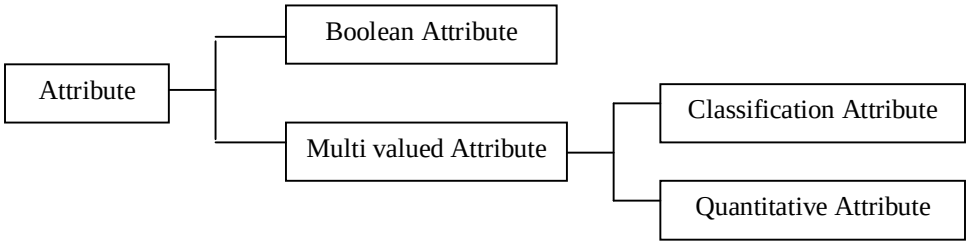

Figure 1 Attribute Classification Chart

Mining association-rule of multiple valued attributes is different from those of Boolean's. At the very beginning, Shapiro described the association-rule of multiple valued attribute as $x=q_x => y=q_y$ which has two problems: One is combinatorial explosion, and the other is that cannot find any practicable association-rule [11-13].

For quantitative attribute, one common method is making the data discretization in data mining, thus translate the association-rule problem of quantitative attribute into Boolean problem. To discretize continuous variable, a problem must concern is interval partition. There are two methods to partition attribute value range, one is partition it into several non-overlapping intervals, then mapping the continuous data to these intervals; the other method is partition into several overlapping intervals. The former method can insure that each attribute case only falls into one interval, but the potential elements nearby borders will be foreclosed possibly by the rigid partition, and cause errors in mining association-rule. Besides, There is not a scientific and reasonable method to determine how to partition interval, in most cases expert experience is a must. As for the latter method, one element may fall into two intervals. The overemphasizing of these elements will affect the validity and effectiveness of association-rule.
The weakness of said two methods is too rigid in border partition. To solve this problem, we can introduce the concept and method of fuzzy mathematics, to consider all probable values of attribute as a Domain. By defining the fuzzy set in attribute domain, the partition border will be mollified. Fuzzy set can ensure that no one element will be excluded, at the same time

however, border element will not be overemphasized. The grade of membership can describe the degree of an element belonging to the set.

For practicable application, fuzzy partition is a natural way. Reference [13] provides an animal database. By mining it, Chan found a rule. In felidae, besides those similar features, the bigger one is tiger, the smaller one is cat, and medium size is leopard. Here, "big" and "small" is fuzzy concept. On the basis of records attributes, fuzzy partition translates the numerical value attribute into fuzzy language attribute, which is called membership degree. This translation method is very much alike our thinking process.

For fuzzy partition problem of extracting attribute based on a sample set, Taiwan scholar Hong and his research team have gotten a satisfactory result. In Reference [8], by adopting some techniques such as sample clustering and decision table, they proposed a heuristic algorithm to get fuzzy partition of each attribute and corresponding membership function. This method is verified by a database of insurance company. By studying samples, they fix on some attributes' membership function such as age and wealth, and then predict premium accordingly (accuracy rate is up to 98.16%). The algorithm is improved in Reference [5]. To Iris classification, the accuracy rate of this algorithm is 96.67%. Other algorithms' are lower than it, such as Dasarathy algorithm is 96.67% and C4 is 93.89% [5].

A FUZZY ASSOCIATION-RULE MINING ALGORITHM

Based on Hong's heuristic algorithm of membership functions and *Apriori* approach [9], we propose a fuzzy mining algorithm to explore association rules from given quantitative transactions. Given the membership function of the business database is known.
The proposed fuzzy mining algorithm is as below. Figure 2 is Flow Chart of the Algorithm

Input: Transaction database *D* $1 \le i \le n$ $1 \le j \le m$ including n records *D(i),* m attributes *Aj*, , membership functions of attribute fuzzy set, predefined Minimum Support $MinSup$ and predefined Minimum Confidence $MinConf$.

Output: A series of association-rules.

Part 1: Seek high frequency item-set L_1 with length 1

STEP 1: Using given membership function set, translate the format of attribute value *Aj* (corresponding attribute of each record $D^{(i)}$ in transaction database D)from $v_j^{(i)}$ $(1 \le i \le n, 1 \le j \le m)$ into $\left(mv_{j1}^{(i)} / M_{j1} + mv_{j2}^{(i)} / M_{j2} + \mathrm{K} + mv_{jk}^{(i)} / M_{jk}\right)$, and then build $(n \times q)$ temporary matrix T to store $mv_{jp}^{(i)}$ value.

Where $k = |A_j|$ is the domain quantity of No. *j* attribute fuzzy partition, M_{jp} is the membership function of attribute A_j in No. *p* domain $1 \le p \le k$, and $mv_{jp}^{(i)}$ is membership degree of $v_j^{(i)}$ in No. *p* fuzzy domain; $q = \sum_{j=1}^{m} k = \sum_{j=1}^{m} |A_j|$ is the sum of fuzzy domain.

STEP 2: For each fuzzy partition domain M_{jp} $(1 \le j \le m, 1 \le p \le k)$ calculate cardinal number by using temporary matrix T:

$$sum_{jp} = \sum_{i=1}^{n} mv_{jp}^{(i)}$$

Compare the sum_{jp} with given $MinSup$, then output fuzzy domains that greater than $MinSup$, put them in high frequency item set L (length is 1).

$$L_1 = \{M_{jp} \mid sum_{jp} \geq MinSup, 1 \leq j \leq m, 1 \leq p \leq k\}$$

STEP 3: Let $t = 1$ t is the quantity of element in current high frequency item set.

Part 2: Seek C_{t+1}

STEP 4: Join two high frequency item sets in L_t, each couple generates a candidate high frequency item set C_{t+1} with length $(t+1)$.

Join Algorithm: bi-cycle high frequency L_t, following the latter two operation till the end of cycle.

Select two high frequency item sets L_t^p and L_t^q describe as $\left(L_{item.1}^p, L_{item.2}^p, \mathsf{K}, L_{item.t-1}^p, L_{item.t}^p\right)$ and $\left(L_{item.1}^q, L_{item.2}^q, \mathsf{K}, L_{item.t-1}^q, L_{item.t}^q\right)$ respectively, order of element in high frequency item set can be changed freely.

In the said two high frequency item sets, when $(t-1)$ elements equal but the No. t elements are not equal, join them and put in C_{t+1}.

Insert into C_{t+1}

Select $L_{item.1}^p, L_{item.2}^p, \mathsf{K}, L_{item.t-1}^p, L_{item.t}^p, {}^{L_{item.t}^q}$

From L_t^p, L_t^q

Where $((L_{item.1}^p = L_{item.1}^q) \wedge (L_{item.2}^p = L_{item.2}^q) \wedge \mathsf{K} \wedge (L_{item.t-1}^p = L_{item.t-1}^q) \wedge (L_{item.t}^p \neq L_{item.t}^q))$

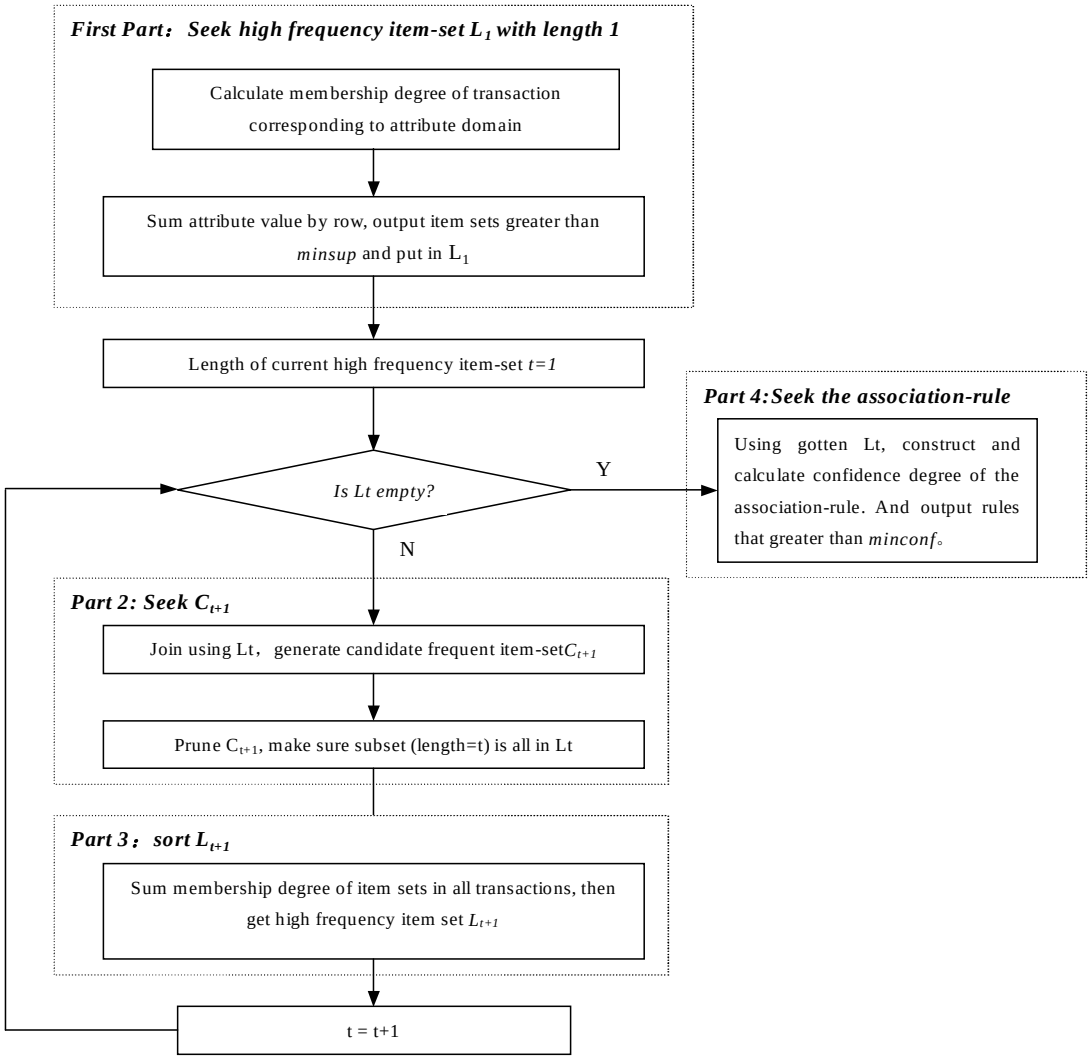

Figure2 Flow chart of the algorithm

STEP 5: Prune candidate high frequency item set C_{t+1}, review each candidate high frequency item set of C_{t+1} and generate a subset with length t, then compare it with L_t to ensure all C_{t+1} subsets appear in L_t, delete unqualified candidate high frequency item sets.

Prune Algorithm: Prune candidate high frequency item sets C_{t+1}, delete candidate high frequency item sets which subset is not in L_t.

Review all candidate high frequency item sets of C_{t+1}, then select a transaction set C_{t+1}^{p} to do following two operations till the end of cycle.

1 Look for all subset of C_{t+1}^{p} with length t;

2 Make sure each subset is in L_t, if not, delete C_{t+1}^{p} from C_{t+1}.

Part 3: Seek L_{t+1}

STEP 6: In candidate high frequency item set C_{t+1}, element e (with length $(t+1)$) can be described as $(e_{item.1}, e_{item.2}, \mathrm{K}, e_{item.t}, e_{item.t+1})$, following latter 2 steps, and put qualified high frequency item sets in high frequency item set L_{t+1} (length $(t+1)$).

(1)Calculate fuzzy membership degree $f_e^{(i)}$ of each transaction $D^{(i)}$ corresponding to candidate high frequency item e in database D.

$$f_e^{(i)} = f_{e_1}^{(i)} \wedge f_{e_2}^{(i)} \wedge \mathrm{K} \wedge f_{e_{t+1}}^{(i)} = \underset{r=1}{\overset{t+1}{Min}} f_{e_r}^{(i)}$$

Where $f_{e_r}^{(i)}$ is the membership degree of No. i transaction $D^{(i)}$ in No. $r(1 \le r \le t+1)$ fuzzy partition domain.

(2)Sum membership degree Sum_e of candidate high frequency item e in all transactions, compare it with $MinSup$, and put items that greater than $MinSup$ in high frequency item set L_{t+1};

$$Sum_e = \sum_{i=1}^{n} f_e^{(i)}$$

$$L_{t+1} = \{e \,|\, Sum_e \ge MinSup\}$$

STEP 7: If L_{t+1} is empty, turn to next step; otherwise, let $t = t+1$, repeat steps from 4 to 6.

Part 4: Seek association-rules

STEP 8: Construct and output association rules.

1 For high frequency item e $(e_{item.1}, e_{item.2}, \mathrm{K}, e_{item.t})$ with length t, calculate it's subset e^s that greater than 1.

2 Construct association rule $e^s \Rightarrow (e - e^s)$ calculate confidence degree α; then output rules which confidence degree is greater than $MinConf$.

$$\alpha\left(e^s \Rightarrow (e - e^s)\right) = Sum_e \,/\, Sum_{e^s}$$

AN EXPERIMENTAL EXAMPLE

Fisher's Iris Data containing 150 training instances [14-15] was used to demonstrate the effectiveness of the proposed algorithm. Execution of the algorithm was done on PC (C++, Microsoft Access database).

The program had to distinguish among three species of Iris flowers: setosa, versicolor and verginica (described with F_1, F_2 and F_3). There were 50 training instances for each class. Each training instance was described by four attributes: sepal length (SL),sepal width(SW), Petal length(PLO)and petal width(PW). The unit for all four of attributes was centimeters. According to [8] to calculate membership function of the database (as figure3, figure4, figure5 and figure6). Attribute of leaf's length was classified into 3 fuzzy subsets, which can be described as PL_0, PL_1 and PL_2. Leaf's width was classified as 4 fuzzy subsets, which can be described as PW_0, PW_1, PW_2 and PW_3.

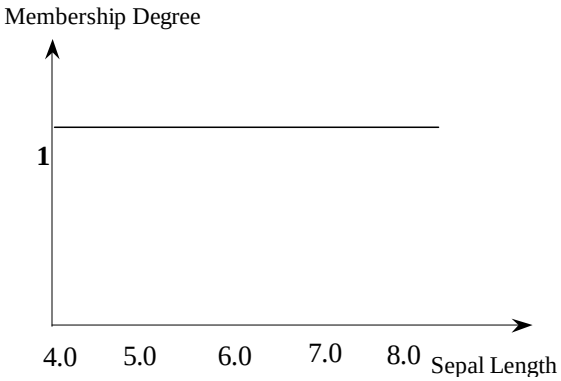

Figure3 Membership Function of Sepal Length

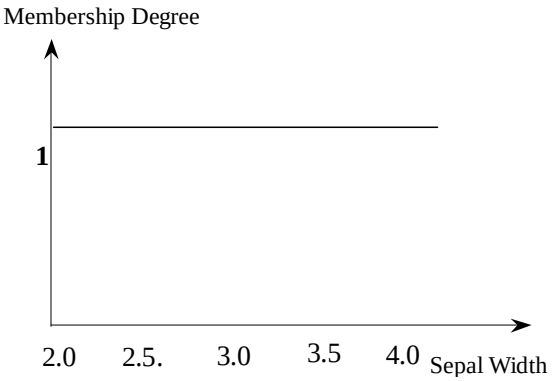

Figure4 Membership Function of Sepal Width

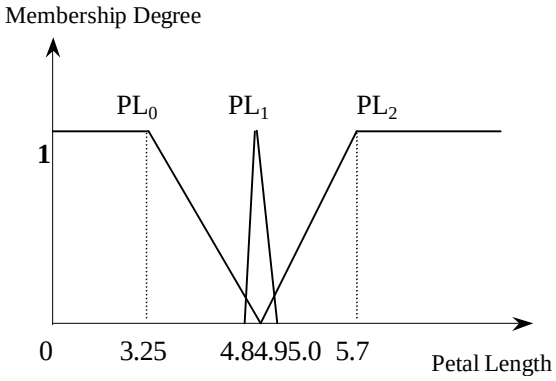

Figure5 Membership Function of Petal Length

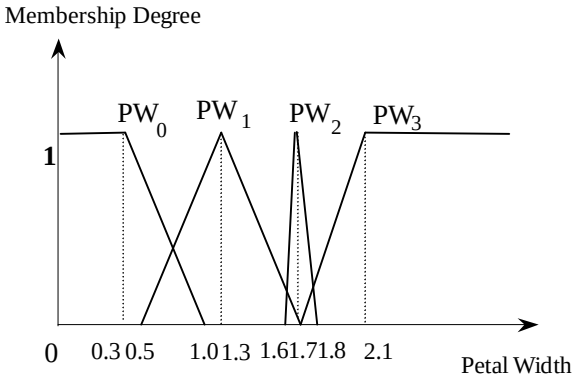

Figure6 Membership Function of Petal Width

EXPERIMENT 1: Input $MinSup = 0.08\%$, $MinConf = 60\%$, and got fifteen association rules as follows:

(1) $PL_0 \wedge PW_0 \Rightarrow F_1$, Confidence=100%;

(2) $PW_0 \wedge PW_1 \Rightarrow F_1$, Confidence=100%;

(3) $PL_0 \wedge PW_1 \Rightarrow F_2$, Confidence=99%;

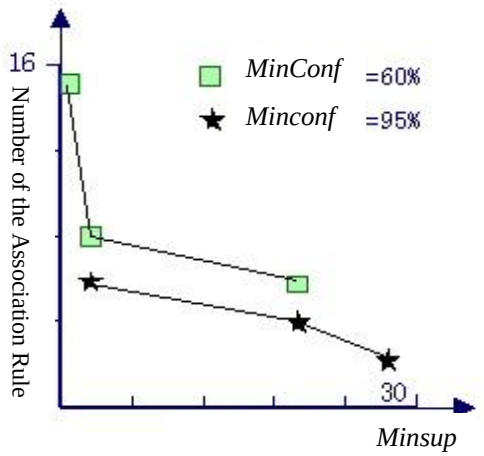

Figure 7 MinConf and Number of the Association Rule

(4) $PL_1 \wedge PW_1 \Rightarrow F_2$, Confidence=100%;

(5) $PL_2 \wedge PW_2 \Rightarrow F_2$,Confidence=100%;

(6) $PL_0 \wedge PW_2 \Rightarrow F_3$,Confidence=100%;

(7) $PL_0 \wedge PW_3 \Rightarrow F_3$,Confidence=67%;

(8) $PL_1 \wedge PW_3 \Rightarrow F_3$, Confidence=100%;

(9) $PL_2 \wedge PW_1 \Rightarrow F_3$, Confidence=85%;

(10) $PL_2 \wedge PW_3 \Rightarrow F_3$, Confidence=100%;

(11) $PL_0 \Rightarrow F_1$, Confidence=69%;

(12) $PW_0 \Rightarrow F_1$, Confidence=100%;

(13) $PW_1 \Rightarrow F_2$, Confidence=94%;

(14) $PL_2 \Rightarrow F_3$, Confidence=77%;

(15) $PW_3 \Rightarrow F_3$, Confidence=99%.

EXPERIMENT 2: Input $MinSup = 1\%$, $MinConf = 60\%$ got eight rules, corresponding to the above rule 1, 3, 10, 11, 12, 13, 14 and 15 respectively, and minimum support is 100%, 99%, 100%, 72%, 100%, 94%, 99% and 99% respectively.

EXPERIMENT 3: Input $MinSup = 1\%, MinConf = 95\%$ got six rules, corresponding to the above rule 1, 3, 10, 12, 14 and 15, and minimum support is 100%, 99%, 100%, 100%, 99% and 99% respectively.

EXPERIMENT 4: Input $MinSup = 20\%, MinConf = 60\%$ got six rules, corresponding to the above rule1, 11, 12, 13, 14 and 15, and minimum support is 100%, 72%, 100%, 94%, 99% and 99% respectively.

EXPERIMENT 5: Input $MinSup = 20\%, MinConf = 95\%$ got four rules, corresponding to the above rule 1, 12, 14 and 15, and minimum support is 100%, 100%, 99% and 99% respectively.

On the basis of the above six experiments and input condition, we can make following conclusion:

(1) The numbers of association rules decreased along with the increase in $MinSup$ and *MinConf*. This is also consistent with our intuition. The small the *MinSup* and *MinConf* values are, the more association-rules can get. Given same *MinConf* value, a lower *MinSup* can get more association-rules. Like Figure 7, the two curves describe the association-rules discovered when $MinConf = 95\%$ and $MinConf = 60\%$ respectively.

(2)Except the above-mentioned data, T.P. Hong and Fisher get eight association rules using a general learning method, corresponding to association rule 1, 3, 4, 6, 7, 8, 9, and 10 of experiment 1. So we can see when *MinSup* and *MinConf* values are appropriate, the algorithm can get all association-rules effectively in practical application.

CONCLUSION AND FUTURE WORK

Based on Hong's heuristic algorithm of membership functions and Apriori approach, we propose a fuzzy mining algorithm to explore association rules from given quantitative transactions. The experimental result shows that this algorithm is effective. The characters of the algorithm are:

1. Introduce fuzzy mathematic into data mining research. By fuzzy partition for multi-valued quantitative attribute, and membership function, translate quantity into fuzzy language. Filtrate and extract high frequency item sets by using concept of membership degree, consequently get association rules. So it introduces a more practical method to mining multi-valued attribute association-rule than Boolean algorithm.

2. Scan transaction database just for one time and save intermediate results in Memory, thereby the running efficiency of the algorithm is improved. But for large and super large database, or instances that have too much intermediate results, it should recur to intermediate database and multiple scanning.

Generally, the algorithm only need scan the database for one time, but if the database is too large, some intermediate results, such as membership degree of attribute fuzzy classification, should save to hare disk, and multiple scanning is also involved. In this condition, one thought is if parallel computation could be a timesaving solution. Moreover, many business databases are distributed storage, and parallel-distributed process will be more effective for solving these complex problems.

ACKNOWLEDGMENTS

This research has been partially supported by a grant from National Natural Science Foundation of China (#70472074). The authors would like to thank the anonymous referees for their very constructive comments.

REFERENCES

1. Agrawal, T. Imeielinksi (1993). A. Swami, Mining association rules between sets of items in large database, *The 1993 ACM SIGMOD Conf.*, Washington, DC, USA 207-216.
2. Agrawal, T. Imeielinksi, A. Swami (1993). Database mining: a performance perspective, *IEEE Trans on Knowledge Data Eeng*., 5,6, 914-925.
3. M.S. Chen, J.han, P.S.Yu(1996). Data mining: an overview from a database perspective, *IEEE Trans on Knowledge Data Eeng*., 8,6, 866-883.
4. A.Famili,W.M. Shen, R.Weber, E.Simoudis(1997). Data Processing and intelligent data analysis. *Intell. Data Anal. 1,1*, 121-132.
5. T. P. Hong, J. B. Chen(1999). Finding relevant attributes and membership functions. *Fuzzy Sets and Systems*, 103, 3,: 389-404.
6. D.G. Burkhardt, P.P.Bonissone(1992), Automated fuzzy knowledge base generation and tuning, in:Proc.IEEE *Internat. Conf. Fuzzy Systems*, San Diego, 179-188.
7. C.C.Lee. Fuzzy logic in control system fuzzy logic controller-Parts I and II,IEEE Trans. *System MAN Cybernet.*
8. T. P. Hong, C. Y. Lee(1996). Induction of membership functions from training examples. *Fuzzy Sets and Systems*, 84,: 33-47.
9. Jiawei Han Micheline Kamber(2001). Data Mining: Concepts and Techniques. *Morgan Kaufmann Publishers*,228.
10. R. Agrawal, H.Mannila, R.Srikant, H.Toivonen, A.I. Verkamo, T. Imeielinksi, A. Swami(1996). Fast discovery of association rules, in:U.M.Fayyad, G.Piatetsky- Shaprio, P.Smyth, R.Uthurusamy(Eds.),*Advances in Knowledge Discovery and DataMining, AAAI Press, Menlo Park,* 307-328.
11. G. Piatetsky Shapiro(1991). Discovery, Analysis, and Presentation of strong rules. In: G Piatetsky Shapiro, W. J. Frawley eds. Knowledge discovery in database. *AAAI/MIT Press*., 229-248.
12. Blake, C.L. & Merz, C.J. (2000). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

13. Keith C. C. Chan, Wai Ho Au(2000). An effective algorithm for mining interesting quantitative association rules. http://www.acm.org. 9-18.

14. R.Fisher (1936). The use of multiple measurements in taxonomic problems, *Ann. Eugenics* 7179-7188.

15. Blake, C.L. & Merz, C.J. (2000). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

16. T.P.Hong, C.S.kou, S.L.Wang(2004). A fyzzy AptioriTid mining algorithm with reduces computational time. *Applied Soft Computing* 5,1-10.

17. T.P.Hong, K.Y.Lin, S.L.Wang(2003), Fuzzy data mining for interesting generalized association rules. *Fuzzy Sets and Systems* 138,255-269